FortressAI & Generative-KI-Sicherheit

Wie FortressAI die Nutzung generativer KI auf Kernel-Ebene steuert – und Datenlecks an ChatGPT, Copilot, Gemini und andere KI-Tools verhindert, bevor sie geschehen.

Was ist FortressAI?

Kurze Antwort: FortressAI ist das generative-KI-Datenschutzprodukt von Cyber Crucible. Es setzt Richtlinien auf der tiefsten Ebene des Betriebssystems durch, überprüft den Datenzugriff auf dem Gerät und blockiert, redigiert oder erlaubt Informationen sicher, bevor sie ChatGPT, Copilot, Gemini oder ein anderes KI-Tool erreichen können.

Das Problem, das es löst

Öffentliche KI-Tools verändern die Art und Weise, wie Menschen arbeiten, und eröffnen gleichzeitig einen neuen Weg, auf dem Daten ein Unternehmen verlassen können. Jeder Prompt birgt ein Risiko. Mitarbeiter teilen häufig sensible Informationen, ohne zu bemerken, dass sie etwas falsch gemacht haben, und herkömmliche Sicherheitstools wurden nie für diesen Zweck entwickelt.

Warum die Durchsetzung auf Kernel-Ebene hier von Bedeutung ist

FortressAI arbeitet auf derselben Tiefe wie der Rest der Cyber-Crucible-Plattform. Das bedeutet, dass es vor *jedem* KI-Tool schützt – einschließlich solcher, die noch gar nicht existieren – anstatt eine Liste bestimmter Anwendungen zu pflegen, die überwacht werden müssen. Es muss ein neues KI-Produkt nicht verstehen, um zu verhindern, dass sensible Daten dorthin gelangen.

Da die Durchsetzung lokal erfolgt, wird der Prompt-Inhalt niemals zur Überprüfung an einen Cloud-Dienst übermittelt. Der Schutz schafft nicht die Gefährdung, die er eigentlich verhindern soll.

FortressAI ist keine separat angebaute Engine. Es ist eine Richtlinienebene über derselben verhaltensbasierten Kernel-Speicheranalyse, die auch den Daten- und Identitätsschutz antreibt – und genau dadurch kann es nicht nur beurteilen, ob ein KI-Tool zugelassen ist, sondern auch, ob es manipuliert wurde.

Was ist Shadow AI, und wer in meiner Organisation erzeugt dieses Risiko?

Kurze Antwort: Shadow AI bezeichnet Mitarbeitende, die KI-Tools nutzen, die von der IT nicht genehmigt wurden oder ihr nicht bekannt sind. Das Risiko beschränkt sich nicht nur auf einfache Mitarbeitende – privilegierte Administratoren und kompromittierte Konten stellen oft die größere Gefährdung dar.

Drei Quellen der Gefährdung

- **Mitarbeitende**, die nicht genehmigte KI-Tools für alltägliche Produktivitätsaufgaben nutzen.
- **Privilegierte Administratoren**, deren weitreichender Zugriff bedeutet, dass alles, was sie einem KI-Tool zuführen, hochsensible Unternehmensdaten enthalten kann.
- **Kompromittierte Konten**, bei denen ein Angreifer mit gültigen Zugangsdaten KI-Tools als Exfiltrationskanal nutzt – Datenverkehr, der wie gewöhnliche Produktivität aussieht.

Dies geschieht bereits

Cyber Crucible hat direkt beobachtet, wie Mitarbeitende mithilfe von KI-Tools massenhaft auf große Datenmengen gleichzeitig zugreifen, um Informationen zu suchen und hochzuladen. Dies ist kein theoretisches, zukünftiges Risiko, das lediglich modelliert wird; es handelt sich um Verhaltensweisen, die bereits heute in realen Umgebungen sichtbar sind.

Die dritte Kategorie verdient besondere Aufmerksamkeit. Ein Angreifer, der KI-Tools nutzt, um Daten abzuziehen, ist kaum von einem engagierten Mitarbeitenden zu unterscheiden – es sei denn, die Durchsetzung erfolgt auf der Ebene des Datenzugriffs, wo Absicht und Richtlinie unabhängig davon bewertet werden können, wer angemeldet ist.

Warum scheitern DLP, KI-Firewalls und private LLMs daran, KI-Datenlecks zu stoppen?

Kurze Antwort: Jeder Ansatz adressiert nur einen Teil des Problems – und das auf der falschen Ebene. DLP kann umgangen werden, KI-Firewalls sitzen am Netzwerkrand und sind blind für Aktivitäten auf dem Gerät, und private LLMs sind teuer, verhindern aber trotzdem nicht, dass Mitarbeiter öffentliche Tools nutzen.

Wo die einzelnen Ansätze zu kurz greifen

- **Private LLMs** sind teuer im Aufbau und Betrieb und verhindern trotzdem nicht, dass jemand ChatGPT in einem Browser-Tab öffnet.
- **KI-Firewalls** prüfen den Netzwerkverkehr, wodurch sie blind sind für das, was auf dem Endpunkt selbst geschieht, sowie für Insider-Aktivitäten.
- **Klassisches DLP** wurde für Dateien und E-Mail-Flüsse entwickelt und wird routinemäßig über Wege umgangen, die es nie überwachen sollte.

Das umgekehrte Logikproblem

Mehrere Ansätze in dieser Kategorie versuchen, Daten vor KI zu schützen, indem sie *eine weitere* KI einsetzen, um die Ausgabe zu filtern. Das erzeugt zusätzliche Kosten und eine neue Abhängigkeit, ohne dort anzusetzen, wo die Daten tatsächlich abfließen – am Endpunkt, im Moment des Zugriffs.

FortressAI setzt auf Kernel-Ebene an, die unterhalb all dieser Schichten liegt, und greift unabhängig davon, welches Tool, welcher Browser oder welche Erweiterung im Spiel ist.

Wie schützt FortressAI Daten in einem Webbrowser?

> **Beta-Funktion.** Die browserbasierte Erkennung und Durchsetzung der KI-Nutzung befindet sich derzeit in der Beta-Phase. Sie ist funktionsfähig und im aktiven Einsatz, sollte jedoch in Ihrer Umgebung evaluiert werden, bevor Sie sich darauf als Kontrollmechanismus verlassen.

Kurze Antwort: FortressAI überwacht die Browseraktivität auf Kernel-Ebene und bewertet sie anhand der Richtlinie. Wenn der Browser, eine Website oder eine Erweiterung versucht, außerhalb der Richtlinie auf Daten zuzugreifen, wird die Berechtigung auf Kernel-Ebene entzogen — und wenn der Browser Anzeichen einer Ausnutzung (Exploitation) zeigt, werden sämtliche Datenrechte entzogen und der Prozess wird angehalten.

Der Ablauf

1. Ein Benutzer öffnet einen Webbrowser — Edge, Chrome oder Firefox.
2. Die Überwachung auf Kernel-Ebene bewertet den Browser sowie jegliche Seiten- oder Erweiterungsaktivität anhand der zugewiesenen FortressAI-Richtlinie.
3. Wenn der Browser, die Website oder die Erweiterung versucht, außerhalb der Richtlinie auf Daten zuzugreifen, wird die Berechtigung auf Kernel-Ebene entzogen.
4. Wenn der Browser Anzeichen einer Ausnutzung zeigt, werden sämtliche Datenrechte entzogen und der Prozess wird angehalten.

Warum der Browser der entscheidende Kontrollpunkt ist

Der Browser ist der Ort, an dem der Großteil der generativen KI-Nutzung tatsächlich stattfindet, und er ist zugleich eine stark angegriffene Angriffsfläche. Cyber Crucible hat dies unmittelbar beobachtet: Seit dem vierten Quartal 2023 sind automatisierte Reaktionen auf Aktivitäten von Chrome, Edge und Chromium von null auf mehrere Tausend gestiegen, wobei anschließend entsprechende CVEs von Google und Microsoft veröffentlicht wurden. Bei einem Unternehmen eskalierte das Muster von einer Handvoll blockierter Angriffe auf nahezu 4.000 über nahezu jede Workstation hinweg.

Woher weiß FortressAI, dass ein KI-Tool nicht manipuliert wurde?

Kurze Antwort: Bevor einem KI-Tool Zugriff auf Daten gewährt wird, prüft FortressAI, ob dieses Tool manipuliert wurde – die Integrität seines Prozesses und der geladenen Bibliotheken wird über dieselbe verhaltensbasierte Speicheranalyse geprüft, auf der auch der Rest der Plattform basiert. Ein zugelassenes Tool, das kompromittiert wurde, erhält Ihre Daten nicht einfach nur deshalb, weil es auf der Zulassungsliste steht.

Genehmigung ist nicht dasselbe wie Vertrauen

Die meisten Ansätze zur KI-Governance enden bei der Zulassungsliste: Ist diese Anwendung erlaubt? Das ist notwendig, aber nicht ausreichend, denn die Identität einer Anwendung und ihre *Integrität* sind unterschiedliche Fragen.

`claude.exe`, ein Copilot-Client, eine ChatGPT-Desktop-App oder ein Browser mit Gemini können alle rechtmäßig genehmigt sein – und alle ausgenutzt werden. Ein Angreifer, der einen genehmigten KI-Client kompromittiert, erbt jeden Datenzugriff, der diesem Client gewährt wurde. Eine Zulassungsliste allein übergibt diesen Zugriff einfach.

FortressAI prüft daher zwei Dinge, bevor Datenzugriff gewährt wird:

1. **Ist dieses Tool autorisiert?** (Bewertung und Richtlinie pro Tool)
2. **Ist dieses Tool noch es selbst?** (wurde der Prozess oder seine Bibliotheken manipuliert)

Beides muss zutreffen.

Was „manipuliert" hier bedeutet

Die Prüfung stützt sich auf Speicheranalysen – dieselbe grundlegende Sensorebene, die auch dem Daten- und Identitätsschutz zugrunde liegt. Sie untersucht den Zustand des laufenden Programms und seiner geladenen Bibliotheken auf Anzeichen von Injektion, Speichermodifikation oder Ausnutzung.

Dies ist wichtig, weil moderne Angriffstechniken gezielt vertrauenswürdige Prozesse ins Visier nehmen. Code, der in eine genehmigte Anwendung eingeschleust wird, erbt deren Berechtigungen und deren Reputation. Die Überprüfung der Datei auf der Festplatte belegt nichts darüber, wozu der Prozess im Arbeitsspeicher geworden ist.

Was bei einer fehlgeschlagenen Prüfung passiert

Die Reaktion folgt demselben abgestuften Modell, das in der gesamten Plattform verwendet wird:

- **Autorisiert und unmanipuliert, im Rahmen der Richtlinie** — Zugriff wird gewährt.
- **Autorisiert und unmanipuliert, aber über die Richtlinie hinausgehend** — je nach Ihren Regeln blockiert, geschwärzt oder mit gefälschten Daten versehen. Das Tool läuft weiter.
- **Manipuliert oder unbekannt** — abgelehnt oder ausgesetzt, gemäß der Verhaltensengine und Ihren Einstellungen.

Die praktische Auswirkung: Ein kompromittierter KI-Client wird wie ein Angreifer behandelt, nicht wie eine genehmigte Anwendung, die gerade einen schlechten Tag hat.

Kann ich kontrollieren, welche KI-Tools mein Unternehmen nutzen darf?

Kurze Antwort: Ja. FortressAI bietet eine Bewertung pro Tool — granulare Kontrolle darüber, welche KI-Tools genau autorisiert und welche blockiert sind — sowie standortbasierte Durchsetzung, sodass Richtlinien an den Ort geknüpft werden können, an dem sensible Daten liegen, statt an einzelne Anwendungen.

Zwei sich ergänzende Kontrollmechanismen

- **Bewertung pro Tool:** Legen Sie gezielt fest, welche KI-Anwendungen (Gemini, Copilot und andere) in Ihrer Umgebung zulässig sind.
- **Standortbasierte Durchsetzung:** Definieren Sie Schutzmaßnahmen rund um die relevanten Datenstandorte — sodass Quellcode im Repository verbleibt und nicht in einem Chatbot landet — wobei die Zuweisungen von den Nutzern gesteuert werden.

Das Ampelprinzip (Traffic Light Protocol)

Die FortressAI-Richtlinie wird als Ampel dargestellt:

- **Rot** — FortressAI blockiert den Zugriff des KI-Tools auf die zugewiesenen Daten.
- **Gelb** — der Zugriff wird bedingt bewertet, einschließlich dynamischer Behandlung sensibler Daten. *(Die bedingten Abläufe, die von der Purview/MIP-Label-Durchsetzung und der automatischen PII/PDPL-Anonymisierung abhängen, befinden sich in der Beta-Phase — siehe „Welche FortressAI-Funktionen sind allgemein verfügbar und welche befinden sich in der Beta-Phase?“)*
- **Grün** — die genehmigte Nutzung erfolgt normal.

Dies ermöglicht es Unternehmen, KI bewusst einzuführen, anstatt zwischen pauschalen Verboten, die umgangen werden, und einem offenen Zugang, der Daten preisgibt, wählen zu müssen.

Welche FortressAI-Funktionen sind allgemein verfügbar und welche befinden sich in der Beta-Phase?

Kurze Antwort: Die Kern-Durchsetzung auf Kernel-Ebene – das Blockieren sensibler Daten vor dem Zugriff durch KI-Tools, die Autorisierung pro Tool und die standortbasierte Datenrichtlinie – ist allgemein verfügbar. Die browserbasierte Erkennung und Durchsetzung der KI-Nutzung, die Durchsetzung von Microsoft Purview/MIP-Vertraulichkeitsbezeichnungen sowie die automatische PII/PDPL-Anonymisierung befinden sich derzeit in der Beta-Phase.

Allgemein verfügbar

- **Datenschutz auf Kernel-Ebene** — sensible Daten werden daran gehindert, über ChatGPT, Copilot, Gemini oder ein beliebiges anderes KI-Tool nach außen zu gelangen; die Durchsetzung erfolgt auf der tiefsten Ebene des Betriebssystems.
- **Bewertung pro Tool** — granulare Kontrolle darüber, welche KI-Tools genau autorisiert und welche blockiert werden.
- **Standortbasierte Durchsetzung** — Richtlinien, die an die relevanten Datenstandorte gebunden sind, sodass Quellcode im Repository verbleibt und nicht in einem Chatbot landet.

In der Beta-Phase

- **Browserbasierte Erkennung und Durchsetzung der KI-Nutzung** — Überwachung und Durchsetzung von Richtlinien in Bezug auf Browser-, Seiten- und Erweiterungsaktivitäten.
- **Durchsetzung von Microsoft Purview / MIP-Vertraulichkeitsbezeichnungen** — Erweiterung der bestehenden Purview-Klassifizierung auf richtlinienbasierte Kontrolle pro KI-Tool, sodass Organisationen, die Daten bereits klassifizieren, kein zweites Schema pflegen müssen.
- **Automatische PII-/PDPL-Anonymisierung** — dynamische Filterung, die personenbezogene Daten aus ansonsten zulässigen Prompts entfernt. Dies unterstützt den von den meisten Organisationen gewünschten Mittelweg: Mitarbeitende können KI

produktiv nutzen, während sensible Elemente automatisch entfernt werden, anstatt sich auf die Selbstzensur der Mitarbeitenden zu verlassen.

Beta-Funktionen sind funktionsfähig und im aktiven Einsatz, sollten jedoch in Ihrer Umgebung evaluiert werden, bevor Sie sich für eine Kontrolle darauf verlassen. Für aktuelle Informationen zur Verfügbarkeit der Beta-Phase und zur Anmeldung wenden Sie sich an Ihren Cyber-Crucible-Ansprechpartner.

Wie unterstützt FortressAI bei Compliance und Audits?

Kurze Antwort: FortressAI verhindert das Datenleck, anstatt es im Nachhinein zu melden, und liefert klare Aufzeichnungen darüber, wer worauf zugegriffen hat und wo Lecks gestoppt wurden — Nachweise, die Sie einem Auditor vorlegen können.

Prävention als Compliance-Haltung

Die meisten Compliance-Risiken bei der Nutzung von KI folgen demselben Muster: Sensible Daten verlassen die Organisation, und daraus ergibt sich die Verpflichtung zur Untersuchung, Benachrichtigung oder Offenlegung. Wird der Zugriff bereits auf Kernel-Ebene blockiert, tritt das Offenlegungsereignis gar nicht erst ein.

Kontrolle nachweisen

Auditoren und Regulierungsbehörden erwarten in der Regel zweierlei: den Nachweis, dass eine Kontrolle existiert, und den Nachweis, dass sie funktioniert. FortressAI unterstützt beides — eine Richtlinie, die technisch auf Betriebssystemebene durchgesetzt wird, sowie klare Aufzeichnungen, die zeigen, wo die Durchsetzung erfolgt ist.

Wo dies zur Anwendung kommt

Diese Anforderung findet sich in vielen Rahmenwerken wieder — HIPAA für Patientendaten, FERPA für Schülerdaten, GDPR für personenbezogene Daten sowie vertragliche Vertraulichkeitspflichten. Die zugrunde liegende Kontrolle ist in jedem Fall dieselbe: Sensible Daten dürfen die Grenze nicht verlassen, und Sie müssen nachweisen können, dass dies auch nicht geschehen ist.